Clustering of Text Data

Sagar Hukkire

20.06.2017

Part I
Kmeans Method

1 Parameters :

- Clusters =5
- tolerance =0.0001
- iterations = 500 for single run
- initialization =k-means++

2 Discussion of Approach

In real world applications of clustering, such as given uncluster_txt doc which has no labels. In order to cluster it into different clusters ,by grouping the samples based on feature similarity, the Kmeans is more suitable. In this project we are able to cluster the given document into different clusters but the score values are not that much good. We tried with different initilization like random,k-mean++ and k-means++ is more promising method to select appropriate cluster centroid.

3 Conclusion.

With Kmeans also there is no significant improvement in result. Deep learning way can be more suitable for this clustering problem.

Part II

With Counting of Frequency word method

4 Approach:

1. Given tags per cluster are assigned to clusters.
2. Then unclustered document is processed and term frequency counter matrix is calculated.
3. With given clusters are checked in features of unclustered text file.

5 Conclusion:

1. This approach is pretty easy , but it is kind of overfitting.
2. In large scale dataset, the number of tags will grow exponentially ,according clustering will be more complex